CANCER INCIDENCE RATES ADJUSTED FOR REPORTING DELAY

Timely and accurate calculation of cancer incidence rates is hampered by **reporting delay**, the time elapsed before a diagnosed cancer case is reported to the NCI. Currently, the NCI allows a standard delay of 22 months between the end of the diagnosis year and the time the cancers are first reported to the NCI in November, almost 2 years later. The data are released to the public in April of the following year. For example, cases diagnosed in 2000 were first reported to the NCI in November 2002 and released to the public in April 2003. However, in each subsequent release of the SEER data, all prior diagnosis years (e.g., diagnosis years 1999 and earlier in the 2002 submission to the NCI) are updated as either new cases are found or new information is received about previously submitted cases. The submissions for the most recent diagnosis year are, in general, about two percent below the number of cancers that will be submitted for that year in the future, although this varies by cancer site and other factors. The idea behind modeling reporting delay is *to adjust the current case count to account for anticipated future corrections (both additions and deletions) to the data*. These adjusted counts and the associated delay model are valuable in more precisely determining current cancer trends, as well as in monitoring the timeliness of data collection -- an important aspect of quality control (Clegg et al., 2002). Reporting delay models have been previously used in the reporting of AIDS cases (Brookmeyer & Damiano, 1989; Pagano et al., 1994; Harris, 1990).

In this report, we show SEER age-adjusted incidence rates and trends, along with their calculated delay adjustments, for all cancers combined (malignant only except for urinary bladder), for female breast in situ, for urinary bladder (in situ and malignant), and for 15 malignant cancer sites: melanoma (for whites only), lung/bronchus, colon/rectum, prostate, female breast, liver and intrahepatic bile duct, pancreas, cervix uteri, ovary, kidney and renal pelvis, brain and other nervous system, non-Hodgkin lymphoma, all leukemias, esophagus, and stomach.

A delay distribution models the probability of a cancer being reported after a delay of d years (d = 2, 3, ...21). The number of cancers reported at each delay year is assumed to follow a Poisson distribution. Cases are removed as corrections to the data are made, and the probability of removing cases is modeled as a binomial distribution. To reduce the number of parameters that have to be estimated and to achieve stability in the tails of the delay distributions, an assumption is made that all cancer cases will be reported within 21 years of diagnosis.

The delay distributions were modeled as a function of covariates using a discrete-time proportional hazards model. For the models presented here the following potential covariates are included: age at diagnosis, sex, diagnosis year, delay times, and race/ethnicity. Age at diagnosis was modeled as a 3-category variable: <50, 50-64, 65+. Diagnosis year was modeled either as a continuous covariate or as categorized variables: 1981-1985, 1986-1990, or 1991-2000. Delay time d was modeled as a categorical variable in one of three ways: (1) $d > 2$ or $d > 3$, (2) $d > 2$, $d > 3$, $d > 4$, or $d > 5$, and (3) $d > 2$, $d > 3$, ... , $d > 10$. Only blacks and whites were analyzed. For melanoma, only whites were analyzed because melanoma is rare for blacks.

Maximum likelihood estimates of delay probabilities were obtained using the Newton-Raphson algorithm. For each of the cancer sites, models of many combinations of covariates were considered. We evaluated the models by fitting the models using data from each of the annual data submissions between 1983 and 2001 and then predicting the counts for the 2002 submission. For each cancer site, the model that minimized the sum of squared prediction errors was chosen as the default final model. However, to choose a more parsimonious model, we added an additional selection step in which possible competing models were selected using the following criteria: (1) the competing model had fewer number of parameters of the default model, and (2) the percent change between the prediction errors of the competing and the default models per extra parameter (i.e., percent change in prediction errors divided by the difference in the numbers of parameters between the two models) was less than 1%. If more than one competing model met the criteria, the model with the smallest percentage change per extra parameter was generally selected. However, if here are other competing models that had fewer parameters and the differences between their percentage changes per extra parameter and the smallest one did not exceed

0.02, the competing model with the fewest number of parameters (rather than the model with the smallest percentage change per extra parameter) was selected. The chosen model was then refitted using all data (1983-2002 submissions, 1981-2000 diagnosis years) to estimate delay distributions and calculate delay-adjusted estimates of the cancer counts.

Age-adjusted (using the 2000 US standard million population) cancer incidence rates were then calculated with and without adjusting for reporting delay. Joinpoint linear regression was used to obtain the annual percentage changes for the 1973-2000 incidence rates for the data series with and without delay adjustment. Because the delay distribution was assumed complete after 21 years, incidence rates for diagnosis years prior to 1981 were not reporting-adjusted. In joinpoint regression analyses, up to three change points (i.e, 4 trend-line segments) were allowed, and these were modeled to fall at either whole years or midway between diagnosis years. Change points were constrained to be at least 2 years away from both the beginning and the end of the data series and at least 2 years apart. Models were fitted using weighted least squares (weighted by appropriate variances of age-adjusted incidence rates) of the joinpoint regression software.

Results show that adjusting for delay tends to raise cancer incidence rates in more current reporting years. While this adjustment increases the rate of change over the most recent diagnosis years, it probably will only rarely cause the detection of a new joinpoint, although this is possible. See Clegg et al. (2002) for details on the impact of reporting-delay adjustment to SEER cancer incidence rates.

References

Clegg LX, Feuer EJ, Midthune D, Fay MP, Hankey BF. Impact of Reporting Delay and Reporting Error on Cancer Incidence Rates and Trends. Journal of the National Cancer Institute 2002;94:1537-1545.

Brookmeyer R, & Damiano A. Statistical methods for short-term projections of AIDS incidence. Statistics in Medicine 1989;8:23-34.

Pagano M, Tu XM, De Gruttola V, & MaWhinney S. Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. Biometrics 1994;50:1203-1214.

Harris JE. Reporting delays and the incidence of AIDS. Journal of the American Statistical Association 1990;85:915-924.